📘 CUSTOMER SATISFACTION PREDICTION – FINAL PROJECT SUMMARY

---

1. Project Overview

This project aims to predict **Customer Satisfaction Ratings** using structured + unstructured customer support ticket data.
In addition, the project delivers:

✓ **Classification Model** – Predict satisfaction (3 classes)
✓ **Regression Model** – Predict ticket resolution time
✓ **Customer Segmentation** – Clustering customers based on behavior
✓ **NLP Analysis** – Extract insights from ticket descriptions
✓ **Explainability** – SHAP-based feature importance
✓ **Deployment-Ready Files** – Models saved in joblib format

The solution is designed for **Customer Support Analytics**, **Product Teams**, and **Business Stakeholders**.

---

2. Dataset Description

The dataset contains **8,320 customer support tickets** with the following key fields:

Customer Information

- Customer Name
- Customer Email
- Customer Age
- Customer Gender

Ticket Information

- Product Purchased
- Ticket Type
- Ticket Subject
- Ticket Description (NLP text)
- Ticket Priority
- Ticket Channel
- First Response Time
- Time to Resolution
- Customer Satisfaction Rating (Target)

Engineered Features

- first_response_hours
- resolution_hours
- word_count
- desc_len
- Response Delay (hrs)

---

3. Business Problem

Companies struggle to understand **why customers are dissatisfied** and **which operational factors drive CTS scores down**.

This project provides:

🎯 Predictive insights
🎯 Strong explainability using SHAP
🎯 Ticket resolution forecasting
🎯 Customer segmentation for personalization

---

4. Preprocessing Pipeline

Cleaning

- Missing numerical values → median imputation
- Missing categorical values → "Unknown"
- Corrected datetime formats
- Removed invalid timestamps
- Extracted numeric time durations

Encoding

- Label Encoding for:
    - Gender
    - Product
    - Ticket Priority
    - Ticket Channel
    - Ticket Type
    - Ticket Status

Text Processing

- TF-IDF vectorization of **Ticket Description**
- Stopword removal
- TF-IDF features reduced using SelectKBest(chi2)

Scaling

- StandardScaler applied on numeric features

---

5. Final Features Used

Categorical Features

['Customer Gender', 'Product Purchased', 'Ticket Type',

'Ticket Status', 'Ticket Priority', 'Ticket Channel']

Numeric Features

['Customer Age', 'Customer Satisfaction Rating',

'Response Delay (hrs)', 'word_count', 'desc_len']

---

6. Model Development

Models Tested

Model	**Type**	**Notes**
Random Forest	Classification	Good stability
XGBoost	Classification	Best for SHAP explainability
SVM (balanced)	Classification	*Best performing model*
RF Regressor	Regression	Best for resolution forecasting
K-Means	Segmentation	k = 4 clusters

---

7. Final Model Scores (Classification)

Model	**Accuracy**	**Macro F1**
RandomForest (weighted)	**0.7999**	**0.6123**
XGBoost (default)	**0.7946**	**0.5989**
SVM (balanced) ⭐	**0.8093**	**0.6326**

🏆 **Best Classification Model: SVM (balanced)**

✓ Best accuracy
✓ Best macro F1
✓ Best balance on minority classes

---

8. Final Tuned Model Scores

XGBoost (Tuned)

Accuracy: 0.7946

Macro F1: 0.5989

SVM (Tuned – Final Production Model)

Accuracy: 0.8093

Macro F1: 0.6326

Classification Report

Accuracy: 0.81

Macro F1: 0.63

Class 0 (Low Satisfaction): F1 = 0.49

Class 1 (Neutral): F1 = 0.95

Class 2 (High Satisfaction): F1 = 0.46

---

9. Regression Model – Resolution Time Prediction

Model Used: RandomForestRegressor

Metric Score

MAE 7.85 hours

MSE 93.29

RMSE 9.66 hours

Interpretation:
🕒 Average prediction error ~ **8 hours**
🕒 Reasonable for operational analytics

---

10. Explainability – SHAP Summary (XGBoost)

Top features influencing satisfaction:

Most Important Drivers

1. **Response Delay (hrs)**
2. **Ticket Priority**
3. **Ticket Type**
4. **Customer Age**
5. **Text Description Word Count**
6. **Sentiment-driven TF-IDF keywords**
7. **Ticket Channel (Email/Chat/Phone)**

Insights:

- Higher response delay → lower satisfaction
- Critical priority tickets → polarizing scores
- Billing & technical issues → lower scores
- Long ticket descriptions → higher frustration

---

11. Customer Segmentation (KMeans, K=4)

Cluster Summary

Segment Description

Cluster 0 Older customers, long descriptions, slow resolution

Cluster 1 Young customers, fast resolution, high satisfaction

Cluster 2 High priority, multi-channel users

Cluster 3 Fast responders, low issue complexity

Segmentation helps personalize support strategies.

---

12. Final Recommendations

A. Use SVM model in production

- Best accuracy and balanced F1
- Handles complex, nonlinear patterns
- Stable across training runs

B. Use XGBoost for Explainability

- Supports SHAP
- Gives clear feature importance

C. Improve Data Quality

- Collect proper resolution timestamps
- Add customer sentiment scores
- Improve ticket type taxonomy

D. Improve Models Further

- Use BERT embeddings for ticket descriptions
- Try CatBoost for mixed data
- Add anomaly detection for escalations

---

13. Deployment Summary

Files Saved

model_artifacts/

svm_model.joblib

xgb_shap_model.joblib

rf_regressor.joblib

tfidf.joblib

scaler.joblib

encoders.joblib

Can Deploy Using

- **Streamlit UI**
- **FastAPI backend**
- **Docker + Cloud Run / EC2**

FINAL EXECUTIVE SUMMARY

This project successfully delivers a complete **Customer Experience AI solution** with:

✓ Customer Satisfaction Prediction
✓ Ticket Resolution Time Forecasting
✓ Customer Segmentation

✓ NLP + Explainability

✓ Deployment-ready components

The **SVM classification model** achieved the **best performance (Accuracy 81%, Macro F1 0.63)** and will be deployed in production.
For interpretability, the **XGBoost model** is used with SHAP to provide transparent insights for business teams.
Regression and segmentation modules allow deeper operational improvements.

This end-to-end solution is ready for **deployment**, **presentation**, and **business decision making**.